From Capturing to Rendering: Volumetric Media Delivery With Six Degrees of Freedom

Jeroen van der Hooft, *Member*, *IEEE*, Maria Torres Vega, *Member*, *IEEE*, Tim Wauters, *Member*, *IEEE*, Christian Timmerer, *Senior Member*, *IEEE*, Ali C. Begen, *Senior Member*, *IEEE*, Filip De Turck, *Senior Member*, *IEEE* and Raimund Schatz, *Member*, *IEEE*

***Abstract*—Technological improvements are rapidly advancing holographic-type content distribution. Significant research efforts have been made to meet the low-latency and high-bandwidth requirements set forward by interactive applications such as remote surgery and virtual reality. Recent research made six degrees of freedom (6DoF) for immersive media possible, where users may both move their head and change their position within a scene. In this article, we present the status and challenges of 6DoF applications based on volumetric media, focusing on the key aspects required to deliver such services. Furthermore, we present results from a subjective study to highlight relevant directions for future research.**

I. Introduction

"Help me, Obi-Wan Kenobi. You're my only hope," said the hologram of Princess Leia in Star Wars: Episode IV - A New Hope (1977). This was the first time in cinematic history that the concept of holographic-type communication was illustrated. Almost five decades later, technological advancements are quickly moving this type of communication from science fiction to reality.

Due to the plethora of applications in the areas of healthcare or Industry 4.0, the three-dimensional representation of objects has received attention in the last years [1]. Research efforts were made to realize six degrees of freedom (6DoF) for immersive media, where the user may both move their head and change their position within a scene. Compared to traditional video, where the user has a passive role, or 360° video, where the user can only turn their head, 6DoF introduces additional complexity in terms of content representation and encoding. Most importantly, because of the user's shift from a rather passive to an active role, real-time interaction with the content becomes crucial. This increases requirements in terms of latency (in the order of tens of milliseconds) and bandwidth (Gbps or even Tbps), making it difficult to keep the user's quality of experience (QoE) at high levels [2].

In this article, we present the status and challenges of 6DoF media delivery from a QoE perspective. Starting from a relevant use case, we propose an architecture for streaming immersive media. This architecture is presented in Section II, providing a high-level overview of the envisioned components. Sections III through VI discuss these components, elaborating on the status and challenges of each. In Section VII, the details and analysis of a subjective evaluation study are provided. Finally, the article is concluded in Section VIII.

II. Envisioned Architecture

We envision an end-to-end system for 6DoF immersive media streaming. To present and discuss the required components, we will target the following use case. Four people want to create a virtual scene in which they are featured, so that the scene can be consumed by interested users. Example scenarios include entertainment (e.g., a band releasing an immersive music video) or educational purposes (e.g., a virtual museum tour). To enable this, the four people need to be captured on camera, so that a three-dimensional, virtual scene can be created in which they reside. Any interested user, whether they are a participant or an audience member, needs to be able to stream the resulting content and, through the use of a head-mounted display (HMD), move freely within the scene. To enable the targeted use case, we propose an architecture that consists of four components (see Figure 1):

- On the source side, objects are captured by multiple cameras. Their output is merged together to create a unified representation of the object, which is processed further to enable delivery over the network;
- The network domain enables the client to stream the content, comprising content delivery networks (CDN) which bring the content closer to the end user and enable scalable delivery. It also provides intelligent means to perform demanding computation tasks.
- The client is responsible for tracking the user's movement and focus, and decides on the video quality at which to download each of the considered objects. Once delivered, the content is decoded and rendered, or used as input for reasoning and control tasks.
- A fourth component enables quality and perception evaluation of the service, *i.e.,* derives the user's QoE. By doing so, different systems and approaches can be compared, identifying those components that require further improvements and research attention.

In the following sections, each of these components will be discussed in detail.

III. The Source Side

A. Content Capture and Representation

To capture three-dimensional objects at different locations, several types of technologies can be considered. Image-based

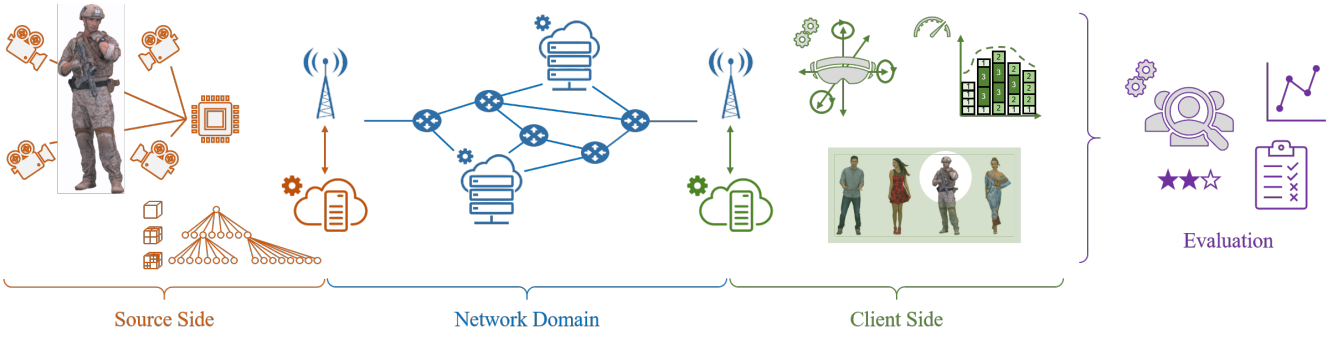

Figure 1. Envisioned architecture for 6DoF immersive media delivery.

Figure 2. The *soldier* object uncompressed (left) and compressed to 12 Mb/s with Mekuria *et al.* [4] (middle) and the V-PCC encoder [1] (right). Reproduced from [1].

solutions require a representation of images at different angles and tilt. Thus, objects are typically captured using a camera array (*i.e.,* a setup with multiple cameras positioned on a grid). Volumetric media-based solutions represent the object as a collection of points (a so-called point cloud), where each of them holds information on the geometry (x, y, z position) and texture (*e.g.,* RGB values). Given the position of each point, the object can be rendered from any viewing angle. The content can be captured through specialized camera setups or by using light detection and ranging (LiDAR)-based cameras. Mesh-based solutions use triangles to represent three-dimensional objects in space. The coordinates of the vertices, along with texture components of the triangles, can be used to render the object based on the user's position and focus.

Meshes better exploit traditional graphics pipelines such as mipmaps or anisotropic filtering, and have been shown to offer higher quality at larget bitrates compared to volumetric media [3]. However, point clouds outperform meshes at lower bitrates, and support dynamic and topology-changing captures in a more straightforward manner. Furthermore, they have the advantage of straighforward tiling and culling compared to meshes, which need to respect triangle and uv parameterization continuity.

The chosen representation has a direct impact on the perceived quality. The amount of memory required by image-based solutions is so high that it is not possible to render high-quality imagery on lightweight commodity hardware. The resulting quality can thus be significantly lower than expected by a consumer used to streaming traditional video. In the case of volumetric media, stitching the input of many different cameras can result in visible artefacts. This is illustrated in the 8i dataset [5], which consists of four moving point cloud objects, each captured during 10 seconds using a frame rate of 30 Hz. Even though the dataset was generated with 42 carefully placed RGB cameras, certain parts of the content are missing as a consequence of inaccurate stitching.

B. Content Encoding

Although volumetric media requires less data than image-based solutions, the amount of storage and transmission resources is still significant. The 8i dataset, for example, consists of four dynamic objects, each with a bitrate between 3.8 and 5.7 Gb/s [5]. Consuming a scene in which all objects are present would require approximately 19.2 Gb/s. Even when considering 5G, such transmission speeds are infeasible. Therefore, compression techniques are necessary.

The codec by Mekuria *et al.* [4] uses the correlation between subsequent point cloud frames to achieve better compression performance. To this end, the bounding box of the point cloud is recursively divided into eight subparts, corresponding to the eight children of a tree-based structure (see Figure 1). Only non-empty children are subdivided further, resulting in a so-called octree of voxels, each of which is represented by the coordinates of its center. Once this subdivision is made for consecutive frames, a transformation is computed using the iterative closest point algorithm, which is then compressed by applying a quaternion quantization scheme [4].

This codec was used as a benchmark in a call for proposals, launched by MPEG in 2017. Out of nine submissions, MPEG selected the reference encoder for video-based point cloud compression (V-PCC) [1]. This codec decomposes the point cloud as a set of patches through an orthogonal projection to a two-dimensional grid. Next, it merges the different patches into two separate video sequences that capture the geometry and texture information, respectively. It then applies traditional video coding techniques to compress both. Overall, it results in a higher visual quality for the same bitrate (see Figure 2), and offers lossy compression ratios between 100 and 500 [1].

While the encoding time is less relevant for video on demand, it strongly contributes to the end-to-end delay in live video. Using V-PCC, non-parallelized encoding of 30 frames of the *soldier* object takes approximately 4 hours on a hexacore

Intel(R) Xeon(R) CPU E5645 @ 2.40 GHz with 24 GB of RAM. To minimize this encoding time, the amount of data to process should be reduced. In the case of sports applications, for example, a health coach might only be interested in the movement of the upper body. A so-called culling process could then remove redundant data, enabling faster processing and delivery. A first research effort in this direction was made in terms of client-side rendering [6], but the concept has not yet been investigated in the case of encoding.

IV. The Network Domain

Once the content has been captured and compressed, it needs to be packaged as streamable units and delivered to the client. Most methods today use HTTP Adaptive Streaming (HAS). In HAS, several representations of different quality and bitrate are provided by a video encoder. The content is divided into segments of one to ten seconds, which are stored on an HTTP-enabled server. The client decides on the quality at which to retrieve each segment, based on the perceived bandwidth, the buffer status and the user's preferences.

A. Application and Transport Layer Optimizations

In traditional video streaming solutions, the lowest streamable unit is a temporal video segment, which can be retrieved by issuing a single GET request using HTTP/1.1 over TCP. When several point cloud objects are considered, however, multiple streamable objects need to be retrieved by the client. Multi-rate encoding can be combined with application layer optimizations such as HTTP/2, which allows the server to push data to the client, thus eliminating the need to send multiple requests. This may reduce the startup delay and help with the latency in interactive delivery. However, the resulting interaction latency due to the built-in congestion and flow control, and the in-order delivery requirement of TCP make it unsuitable for near real-time communication. For this reason, recent solutions revert to UDP-based protocols.

One such example is WebRTC, a suite of real-time communication protocols which has shown promising results for traditional video [7]. WebRTC is, however, peer-to-peer in nature, and thus requires multiple encoders for each peering connection, hampering scalability. A second example is the yet to be standardized HTTP/3 protocol, which is based on QUIC [8]. This protocol establishes a number of multiplexed UDP connections, resulting in independent delivery of multiple streams of data. While HTTP/2 uses a single TCP connection and may experience delays due to head-of-line blocking, HTTP/3 does not have this problem. Google reported a rebuffering rate reduction of 18.0% for desktop users when using QUIC for YouTube, although more recent research showed lower levels of improvement [8].

Although these protocols offer advantages in terms of latency, the absence of adequate control mechanisms can be dramatic for the high bandwidth transmission of 6DoF content. Palmer *et al.* address this issue for traditional video streaming by jointly optimizing the application and transport layer, using an adapted version of the QUIC protocol to reliably deliver key frames, while retrieving others without guarantees [9]. In this way, videos can be streamed at a higher visual quality and with less rebuffering events. The same principle could be applied to volumetric media, with prioritization not only given to certain key frames, but also to the most important spatial regions and objects that make up a scene. The combination of reliable and unreliable delivery could then be used to retrieve different objects and quality representations in a timely manner. Further research is required to determine the applicability of this approach to immersive media use cases.

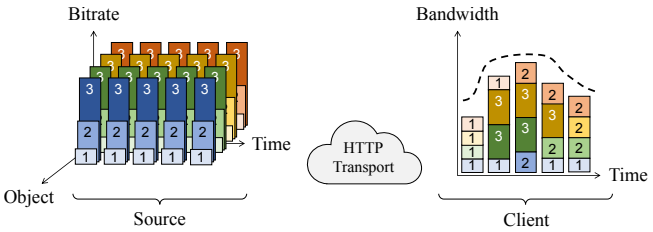

Figure 3. HAS applied to point clouds objects, each encoded using three quality representations (higher is better).

B. Intelligent Network Components

Apart from delivering the content to the client, the network also offers intelligent means to perform demanding computation tasks. This can be achieved through multi-access edge computing (MEC), which enables the devices to access cloud/fog resources on demand. MEC also allows for strategic content caching at the edge, which will play an important role in 6DoF content streaming; upon a new task request, the server/network needs to swiftly decide if it should store the content for future requests or not. Proactive caching strategies need to evolve as well, as they depend on spatiotemporal traffic predictions, the users' location, mobility, *etc.* Other network level approaches such as network coding and network slicing can be exploited to meet the requirements of 6DoF video streaming. The importance of these paradigms is reflected in extensions to the Dynamic Adaptive Streaming over HTTP (DASH) standard, in the form of Server and Network-Assisted DASH (SAND)[1]. SAND allows to share real-time information concerning networks, servers, proxies, cashes and clients with the end user's device, helping the client make better decisions regarding rate adaptation and request the content from the most promising locations first (*e.g.*, from a nearby cache). This principle can be applied to 6DoF media delivery, in that SAND allows to significantly reduce latency when nearby proxies and caches can be used, while the video quality can be improved and QoE-related metrics (see Section VI) can be shared in real time. Further optimizations will become increasingly more important with the advent of 5G and softwarized networking, and should thus gain more research attention in the future.

V. The Client Side

A. Rate Adaptation

Similar to traditional HAS, the client is responsible for deciding on the quality at which to retrieve the content. When

[1]https://www.iso.org/standard/69079.html, accessed 23 July 2020

volumetric media scenes are considered, however, there is also a need for spatial segmentation. Indeed, recent works consider segmentation at the object level (*i.e.,* each object is requested as a whole) [2] and at the level of voxels (*i.e.,* each object consists of multiple streamable units) [10]. Figure 3 shows an example of the former approach, in which the quality of the considered objects goes up with the available bandwidth (the dotted line). It is the responsibility of the client to decide upon the quality of each of the objects, by using an appropriate rate adaptation heuristic. While traditional video streaming solutions take into account properties such as the available bandwidth, video bitrate and buffer status to decide upon the quality of the next video segment, solutions for immersive media also need to take into account the bitrates and locations of the objects within the scene, the position and focus of the user, and the impact of spatial quality differentiation. A preliminary study on rate adaptation heuristics for volumetric media has recently been conducted [2], but more research is needed to unlock their full potential.

B. Viewport Prediction

To avoid video playout freezes, a buffer is generally used at the client side. In order to change the video quality as soon as a user changes their focus, this buffer is kept as small as possible. Still, the rate adaptation heuristic requires accurate information on the user's position and focus in order to allocate the available bandwidth to the most important regions within the scene. For this reason, the so-called viewport prediction is of the utmost importance to 6DoF video streaming: if the user's position and focus can be accurately predicted, the client can compensate for the user's future movement when buffering new content.

Several approaches have been proposed for 360° video, using either content-aware or content-agnostic viewport prediction. In the former case, information on the user's position and focus is used along with information on the video content, by extracting relevant objects and regions through saliency mapping. However, these approaches are time- and resource-demanding, and cannot be applied to videos that have not been processed yet (*e.g.,* in live video scenarios). Therefore, in the latter case, predictions are made based on movement alone, using techniques such as linear regression and neural networks to predict future actions. Compared to 360° video, volumetric media increases the complexity by introducing three additional degrees of freedom. Not only are more advanced approaches needed to predict movement in all directions, saliency mapping also becomes significantly more complex. First steps to predict the user's movement in 6DoF video streaming have recently been made, applying a simple regression model on each of the six degrees of freedom [11]. This approach assumes independence between the different dimensions, a simplified view that results in significant prediction errors, even when predicting movement in the next $100\,\mathrm{ms}$. Thus, further research is required to advance the state of the art forward.

Figure 4. A static scene rendered through the Unity framework[3].

C. Content Decoding

The choice of encoder not only has an impact on the visual quality, but on the decoding step as well. Although the V-PCC encoder outperforms the benchmark encoder of Mekuria *et al.* [4] in terms of visual quality, this technique cannot be used in real time on commodity hardware [2]. The benchmark encoder, however, has recently been applied in a near-real-time volumetric media streaming setup, through numerous optimizations combined with parallel execution on sophisticated hardware [12]. If we plan to use these types of encoders on commodity hardware, such as HMDs or lightweight smartphones, further optimizations are required (*e.g.,* leveraging existing hardware codecs in the experimental V-PCC implementation by Nokia[2]). Alternatively, decoding tasks could be partially offloaded to the network. In the context of 360° video, some works propose to use MEC resources to decode the video content faster [13]. These early works, however, do not consider the placement and timely execution of the required network components. Deciding on where to run each processing task is of great importance to meet the stringent low-latency requirements put forward by immersive media applications, and should be further investigated.

D. Content Rendering

Volumetric media is often rendered through Unity and Unreal Engine (see Figure 4). Once the required data is loaded, users can benefit from the tracking capabilities of HMDs, allowing free movement within the scene. However, non-parallelized loading of 30 frames of the V-PCC decoded *soldier* object requires on average 60.0 seconds on a hexacore Intel(R) Core(TM) i7-8850H CPU @ $2.60\,\mathrm{GHz}$ with $16\,\mathrm{GB}$ of RAM, effectively rendering frames at $0.5\,\mathrm{FPS}$. Thus, a minimum of five CPUs would be required to load and render this *single* object at a target rate of $30\,\mathrm{FPS}$, assuming full parallelization. For this reason, some works again propose to use MEC resources in order to facilitate computational efforts, rendering the user's field of view in the network rather than at the client side. This approach comes with faster execution, but requires real-time communication with the client: information

[2]https://github.com/nokiatech/vpcc, accessed 23 July 2020

[3]https://blog.codecentric.de/en/2020/03/converting-massive-point-clouds-into-vr-scenes-under-unity-virtual-interaction-room-part-10/, accessed 23 July 2020

on the user's position and focus, monitored by the HMD, needs to be transmitted and processed in a timely manner. Some related work already exists for 360° video [13], but fundamental research is required to fully reap the benefits of remote rendering for 6DoF applications.

VI. Objective and Subjective Evaluation

Each of the aforementioned components has an impact on how the system performs. In this regard, a distinction must be made between the quality of service (QoS) and the QoE: the former is objectively measurable and can be retrieved based on the video stream's information and the movement of the user, while the latter is based on subjective sensations and thus requires significant and time-demanding user testing.

Coping with the dynamics of 6DoF video streaming will, however, require near-real-time measurements of how the user perceives the experience. Thus, current research aims to devise models to estimate the QoE of the end user. To assess the visual quality of the rendered field of view, well-known metrics for traditional video streaming have recently been applied [2]. Other examples include the point-to-point and the point-to-plane geometry distortion metrics proposed in MPEG's call for proposals [1]. Measuring the quality alone, however, is not enough: important factors such as the system's latency to react to changes in the user's position, the occurrence of rebuffering events and the startup time need to be taken into account as well. While a significant amount of research on QoE models exists for traditional video, this is not yet the case for immersive media.

Recently, some works attempted to subjectively rate volumetric media, using a passive evaluation protocol to assess the quality degradation of static models due to encoding (*e.g.,* [14]). However, these protocols do not consider the impact of network transport on the perceived quality, nor do they take into account the effects of user interaction. Furthermore, subjective evaluations are mainly performed through double-stimulus tests, in which subjects are asked to rate the degradation of the video compared to the unimpaired source. Although relevant, these approaches do not allow to assess the service's overall quality as perceived by the user. More research is needed to fully understand the user's QoE for 6DoF video streaming solutions.

VII. Analysis

To evaluate the impact of the aforementioned components on the user's QoE, we conducted several subjective experiments in a recent study [15]. Subjects were passively shown a number of source videos between 18 and 24 seconds of length, containing the generated viewport of a scene consisting of four point cloud objects from the 8i dataset [5]. Different movement paths, defined by samples of the user's position and focus, were programmatically defined, *e.g.,* by moving on a straight line or in a circle, zooming in on objects, *etc.*

The considered objects were encoded using the V-PCC encoder with five reference quality representations, each between $2.4\,\mathrm{Mb/s}$ and $53.5\,\mathrm{Mb/s}$ [2]. The objects were made available on a server in an emulated network, so that they could

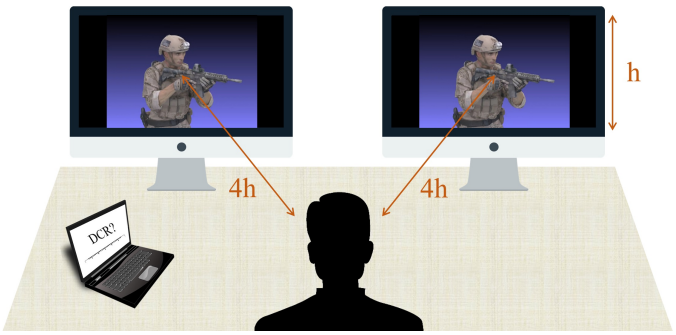

Figure 5. Evaluation setup. The user can rate quality (ACR) or quality differences (DCR) on a continuous scale.

be requested by a headless Python-based client. A buffer size of four seconds was considered, with segments of one second or 30 frames for each object. Traffic control (tc) was used to control the available bandwidth between client and server, which was fixed to 60 Mb/s. Latency was fixed to $37\,\mathrm{ms}$ as a reference value for 4G networks, while no additional packet loss was introduced.

In the original study, only single-stimulus experiments were conducted. Users were asked to rate the visual quality of the presented stimulus using the absolute category rating (ACR). In this paper, double-stimulus experiments were added, in which the uncompressed point cloud objects were shown on one screen and the streamed content was shown on the other (see Figure 5). Subjects were asked to evaluate the difference in quality between the two screens using the degradation category rating (DCR). A continuous 7-point scale[4] was used in both experiments.

A total of 60 subjects participated in our subjective experiments; half of the subjects participated in the double-stimulus study, while the others participated in the single-stimulus study. We eliminated four subjects from the first and two subjects from the second study, as they did not pass an outlier screening procedure. This results in subject counts of $N_1 = 26$ and $N_2 = 28$ for the remaining analysis. Here, we consider three types of videos for each source clip:

- Compressed point clouds are used to render the field of view, respecting a total file size of $60\,\mathrm{Mb}$ for every second of content. The most recent (MR) information on the user's position and focus is used to buffer new point cloud segments that will be consumed in the near future;
- Similar to the above, but content is buffered using perfect knowledge of what the user will focus on in the near future, *i.e.,* assuming clairvoyant prediction (CV).
- Uncompressed content is used (hidden reference, ∞);

Figure 6 shows the cumulative distributions of the opinion scores obtained in our experiments. The double-stimulus results (red) show that some users did not pick up on the hidden reference, assigning scores between 27 and 100. Overall, however, results are significantly better than those for scenarios in which a limited bandwidth ($60\,\mathrm{Mb/s}$) is available. Thus, compression has a negative impact on the QoE and better compression schemes for volumetric video are desirable.

[4]https://www.itu.int/rec/T-REC-P.851-200311-I/en, accessed 23 July 2020

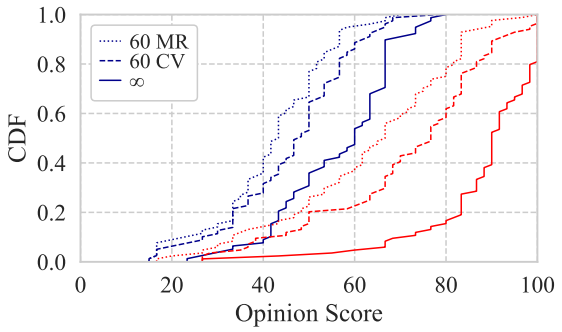

Figure 6. Distribution of the opinion scores for three different scenarios, using single stimulus (blue) and double stimulus testing (red).

Comparing results with and without prediction of the user's movement and focus, the former results in the highest rating. This was expected, since accurate prediction allows the client to buffer a higher quality representation for those objects that are within the field of view at the time of consuming the content. Furthermore, it allows to avoid visual quality switches when the user's attention is shifted from one object to another. This confirms the need for accurate prediction of the user's position and focus in 6DoF applications.

Results also show the major difference between the two tasks the subjects were set to do. Indeed, scores for the single-stimulus experiment (blue) are significantly lower than those for the double-stimulus experiment. Although the general trends between the two approaches are similar, care must be taken to interpret results from any of the two tests. Further research is required to evaluate the impact of the type of study on results for 6DoF immersive media.

Even when the uncompressed point cloud objects were presented to the subjects, the video was rated on the lower end of the spectrum. This aspect was brought up several times during a debriefing interview, with subjects indicating that they had expected a higher visual quality based on their experience with traditional video streaming solutions. This shows that the considered content did not meet the subjects' expectations. Technology to capture three-dimensional scenes thus needs to be improved, providing 6DoF video at a higher visual quality.

We also evaluated the decoding and rendering time using the V-PCC decoder and MPEG's point cloud player on a hexacore Intel(R) Core(TM) i7-8850H CPU @ 2.60 GHz with 16 GB of RAM, considering the source video that focuses on the *soldier* and the *longdress* objects. Two approaches are considered: one in which all objects are decoded and loaded, and one in which only objects that are visible are processed. Using the latter approach, the decoding and rendering time can be reduced by 69.2% and 70.3%, respectively. Even so, the total time required to generate the content is still far from real time, again indicating a need for further research.

VIII. Conclusions

In this article, we presented an overview and challenges of immersive media streaming with six degrees of freedom. Our envisioned architecture has been outlined, consisting of four components located in the source, network and client. Furthermore, we performed a subjective analysis of rendered immersive video, asking participants to rate streaming sessions both by using a double-stimulus and a single-stimulus approach. The results identify several directions and opportunities for further research in this area. In future work, we plan to address these items, focusing both on video streaming as well as on real-time communications.

Acknowledgments

This research is partially funded by Huawei Technologies, China, and by the Christian Doppler Laboratory ATHENA (https://athena.itec.aau.at/). Maria Torres Vega is funded by the Research Foundation Flanders (FWO).

References

[1] S. Schwarz *et al.*, "Emerging MPEG standards for point cloud compression," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 133–148, 2018.
[2] J. van der Hooft *et al.*, "Towards 6DoF HTTP adaptive streaming through point cloud compression," in *ACM Multimedia*, 2019.
[3] E. Zerman *et al.*, "Textured mesh vs coloured point cloud: a subjective study for volumetric video compression," in *IEEE QoMEX*, 2020.
[4] R. Mekuria *et al.*, "Design, implementation, and evaluation of a point cloud codec for tele-immersive video," *IEEE TCSVT*, vol. 27, no. 4, pp. 828–842, 2017.
[5] E. d'Eon *et al.*, "ISO/IEC JTC1/SC29 joint WG11/WG1 (MPEG/JPEG) input document WG1M40059/WG1M74006. 8i voxelized full bodies - a voxelized point cloud dataset," 2017.
[6] S. Discher *et al.*, "Concepts and techniques for web-based visualization and processing of massive 3D point clouds with semantics," *Elsevier Graphical Models*, vol. 104, p. 101036, 2019.
[7] S. Petrangeli *et al.*, "A scalable WebRTC-based framework for remote video collaboration applications," *Springer Multimedia Tools and Applications*, vol. 78, no. 6, p. 7419–7452, 2019.
[8] M. Seufert *et al.*, "QUICker or not? - an empirical analysis of QUIC vs TCP for video streaming QoE provisioning," in *IEEE ICIN*, 2019.
[9] M. Palmer *et al.*, "The QUIC Fix for Optimal Video Streaming," in *ACM EPIQ*, 2018.
[10] J. Park *et al.*, "Rate-utility optimized streaming of volumetric media for augmented reality," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 149–162, 2019.
[11] S. Gül *et al.*, "Low latency volumetric video edge cloud streaming," *arXiv e-prints*, p. arXiv:2001.06466, 2020.
[12] S. Dijkstra-Soudarissanane *et al.*, "Multi-sensor capture and network processing for virtual reality conferencing," in *ACM MMSys*, 2019.
[13] M. S. Elbamby *et al.*, "Toward low-latency and ultra-reliable virtual reality," *Network*, vol. 32, no. 2, pp. 78–84, 2018.
[14] E. Alexiou *et al.*, "A comprehensive study of the rate-distortion performance in mpeg point cloud compression," *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e27, 2019.
[15] J. van der Hooft *et al.*, "Objective and subjective QoE evaluation for adaptive point cloud streaming," in *IEEE QoMEX*, 2020.

Biographies

Jeroen van der Hooft received his M.Sc. and Ph.D. degrees in computer science engineering from Ghent University in 2014 and 2019, respectively. He is currently a postdoctoral researcher at Ghent University - imec, focusing on end-to-end optimizations for adaptive video streaming and low-latency delivery of immersive media.

Maria Torres Vega obtained her M.Sc. and Ph.D. degrees in telecommunication engineering from the Polytechnic University of Madrid and Eindhoven University of Technology in 2009 and 2017, respectively. She is currently a postdoctoral researcher at Ghent University - imec, focusing on the quality of experience in immersive multimedia systems and autonomous management of future networks.

Tim Wauters received his M.Sc. and Ph.D. degrees in electrotechnical engineering from Ghent University in 2001 and 2007,

respectively. He is currently a postdoctoral fellow at Ghent University - imec, focusing on the design and management of networked services for multimedia distribution, cybersecurity, big data and smart cities.

Christian Timmerer received his M.Sc. and Ph.D. degrees from University of Klagenfurt in 2003 and 2006, respectively. He is currently an associate professor at University of Klagenfurt and the chief innovation officer at Bitmovin, focusing on immersive multimedia communication, streaming, and quality of experience. Further details at http://timmerer.com.

Ali C. Begen received his Ph.D. degree in electrical and computer engineering from Georgia Tech in 2006. He is currently a computer science professor at Özyeğin University, focusing on networking and multimedia solutions. Previously, he was a research and development engineer at Cisco. Further details at http://ali.begen.net.

Filip De Turck received his M.Sc. and Ph.D. degrees in electronic engineering from Ghent University in 1997 and 2002, respectively. He is currently a full professor at Ghent University - imec, leading the network and service management research group. His research interests include network and service management, and the design of efficient virtualized networks.

Raimund Schatz received his M.Sc. degree in telematics from TU Graz and his Ph.D. degree in computer science from TU Vienna, while also holding an MBA and M.Sc. degree from Open University. He currently heads the data-driven experience research team at Austrian Institute of Technology and is a postdoctoral researcher at University of Klagenfurt, focusing on quality of experience, pervasive computing and network performance assessment. Further details at http://www.schatz.cc.